iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XI **Month of publication:** November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47284

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

Review on IDS based on ML Algorithms

Om Ambavkar[1], Prathmesh Bharti[2], Amit Chaurasiya[3], Roshan Chauhan[4], Mahalaxmi Palinje[5]
[1, 2, 3, 4, 5]*Electronics & Telecommunication Department, ACE, Mumbai University, Malad (W), Mumbai, India*

Abstract: ***Intrusion detection is one of the challenging problems encountered by the modern network security industry. The developing pace of digital assaults on framework networks as of late compounds the protection and security of PC foundation and PCs. Intrusion Detection and Prevention systems are transforming into a critical part of PC organizations and network safety. Various approaches have been proposed to determine the most effective features and hence enhance the efficiency of intrusion detection systems, the methods include, machine learning-based (ML), Bayesian based algorithm, Random Forest, SVM, Decision Tree. This paper presents an intensive survey on different examination articles that utilized single, hybrid and ensemble classification algorithms. The outcomes measurements, weaknesses and datasets involved by the concentrated on articles in the advancement of IDS were looked at. A future heading for potential explores is likewise given. The paper addressed latest research papers written from the use of machine learning classifiers in intrusion detection systems.***
Keywords: Machine learning, Intrusion Detection System, network, misuse detection, Random Forest

I. INTRODUCTION

Intrusion Detection is the issue of distinguishing unapproved use, abuse, and maltreatment of PC frameworks by both system insiders and outer intruders. Most of the existing commercial IDS products are signature-based but not adaptive or self-learning. As malignant interruptions are a developing issue, we want an answer for distinguish the intrusion precisely. Many techniques were underway to detect the anomalies but had less success. For detecting illicit or abnormal behaviour, IDS is used. Attack is launched in a network in a state of an anomaly behaviour. Attackers use the opportunity of network weaknesses like poor security measures and practices, program bugs such as buffer overflows, yielding the breaches of the network. The rampant usage of internet makes it difficult to protect network resources from the mischievous action of attackers. According to Cybersecurity ventures, the damage related to cybersecurity is predicted to reach $6 trillion yearly by 2021. Gartner reports that, in taking steps to counter the damage, Global expenses on cybersecurity could reaches $133.7 billion in 2022. Multiple measures have been taken in which various security tools such as IDS were developed. [1][3][5]

The system which are utilized for recognition of interruption can be extensively classified into two unique types NIDS and HIDS which represents Network based and host based interruption identification system separately. NIDS are decisively positioned at hubs inside the organization to such an extent that they can play out an examination of the passing traffic on a whole subnet and coordinate it with its own library of predefined assaults. On sensing of an abnormal behaviour of the network or on revelation of an attack, an alert is sent to the administrator. Running against the norm, HIDS runs just on individual hosts or network gadgets. It plays out the checking of the inbound and outbound packets from the device and sends an admonition to the head on detecting any dubious packets. NIDS are of comprehensively two sorts: - anomaly based and signature based. A signature based system is predefined for a specific weakness, so it has a decreased number of misleading up-sides, in this way offering less adaptability. Though, an anomaly based system is more unique in nature and will look for possible attacks after that are out of the predefined ones, thus bringing about a more noteworthy number of false positives. It can identify attacks after just without perceiving the exact kind of assault. [1] Most IDS were developed and evaluated using outdated and old dataset like KDD Cup 99 which lack the most recent and up to date attack labels. Slow detection rate is experienced in the existing works. This happens due to inability to get rid of all redundant and irrelevant columns. High false positive rate. This happens when a legit traffic is incorrectly detected and classified as an attack. The false positive rate increases complexity of IDS, hence, reducing its performance. [2]

II. IDS

Intrusion Detection System (IDS) is an observing system that identifies dubious exercises and produces cautions when they are recognized. Any intrusion action or infringement is regularly revealed either to an overseer or gathered midway utilizing a security data and occasion the event management system [1].

1) *Network Intrusion Detection System:* Network intrusion detection systems (NIDS) are set up at a planned point within the network to examine traffic from all devices on the network. It performs an observation of passing traffic on the entire subnet and matches the traffic that is passed on the subnets to the collection of known attacks. [1]

2) *Host Intrusion Detection System:* Host intrusion detection systems (HIDS) run on independent hosts or devices on the network. A HIDS screens the incoming and outgoing packets from the device just and will caution the executive in the event that dubious or malignant movement is distinguished. It takes a depiction of existing framework documents and contrasts it and the past preview. [2]
3) *Protocol-Based Intrusion Detection System:* Protocol-based intrusion detection system (PIDS) comprises a system or agent that would consistently resides at the front end of a server, controlling and interpreting the protocol between a user/device and the server. It is attempting to get the web server by routinely checking the HTTPS convention stream and acknowledge the connected HTTP convention. [6]
4) *Application Protocol Based Intrusion Detection System:* Application Protocol-based Intrusion Detection System (APIDS) is a system or agent that generally resides within a group of servers It distinguishes the intrusions by checking and deciphering the correspondence on application-explicit protocols. [3]
5) *Hybrid Intrusion Detection System:* Crossover interruption location framework is made by the mix of at least two methodologies of the interruption discovery system. In the hybrid intrusion detection system, have specialist or framework information is joined with network data to foster a total perspective on the organization framework. [4]

III. DETECTION METHOD FOR IDS

1) *Signature based Method:* Signature-based IDS recognizes the assaults with respect to the premise of the particular examples, for example, number of bytes or number of 1's or number of 0's in the network traffic. It additionally recognizes based on the definitely realized vindictive guidance arrangement that is utilized by the malware. The recognized examples in the IDS are known as signatures. [3]
2) *Anomaly-based Method:* Anomaly-based IDS was acquainted with distinguish obscure malware assaults as new malware are grown quickly. In anomaly based IDS there is utilization of AI to make a trustful action model and anything coming is contrasted and that model and it is proclaimed dubious on the off chance that it isn't seen as in model.[3]

IV. ALGORITHMS

1) *Support Vector Machine:* Support Vector Machine or SVM is one of the most famous Supervised Learning algorithms, which is used for Classification as well as Regression issues. Nonetheless, fundamentally, it is utilized for Classification issues in Machine Learning. The objective of the SVM calculation is to make the best line or decision boundary that can isolate n-layered space into classes so we can undoubtedly put the new data of interest in the right classification later on. This best decision boundary is known as a hyperplane. [2]
2) *Decision Tree Algorithm:* Decision Tree is a Supervised learning technique that can be used for both Classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-organized classifier, where internal nodes address the features of a dataset, branches address the decision rules and each leaf node addresses the result. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are utilized to settle on any choice and have numerous branches, while Leaf nodes are the result of those choices and contain no further branches. The decisions or the test are performed based on elements of the given dataset. It is a graphical portrayal for getting every one of the potential answers for an issue/choice in light of given conditions. It is known as a decision tree on the grounds that, like a tree, it begins with the root node, which develops further branches and builds a tree-like design. To fabricate a tree, we utilize the CART algorithm, which represents Classification and Regression Tree algorithm. [5]
3) *Radom Forest Algorithm:* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It very well may be utilized for both Classification and Regression issues in ML. It depends on the idea of outfit realizing, which is a course of consolidating different classifiers to take care of a complicated issue and to improve on the performance of the model. As the name recommends, "Random Forest is a classifier that contains various decision trees on different subsets of the given dataset and takes the normal to work on the prescient precision of that dataset". Rather than depending on one decision tree, the random forest takes the prediction from each tree and in light of the greater part votes of expectations, and it predicts the last result. The more prominent number of trees in the forest prompts higher precision and prevents the issue of overfitting. It requires less preparation investment when contrasted with different algorithms. It predicts yield with high precision, in any event, for the huge dataset it runs effectively. It can likewise keep up with precision when a huge extent of information is absent. [8][9][7] dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing.

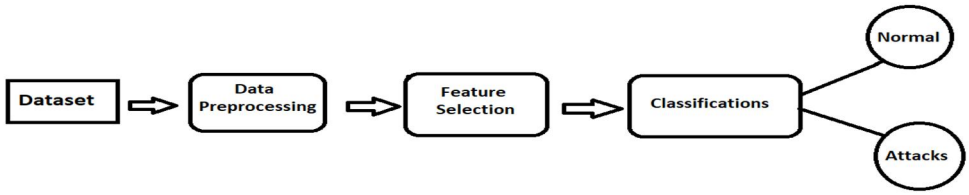
V. OVERVIEW OF IDS USING MACHINE LEARNING

The Machine Learning model focuses on training data sets to predict various class labels. ML is typically divided into three groups:

1) *Supervised Learning:* In supervised ML the dataset to be trained is made up of examples of the input vector, each with their equivalent desired output vectors. Algorithms in this type of learning include: Naïve Bayes, KNN, ANN, Decision Tree SVM, Ensemble methods (Bagging, Voting Classifier), logistic regression. [1][3][2]
2) *Unsupervised Learning:* In unsupervised ML, the learning algorithm is not given labels and as such it must by itself find structure in its input. This is also known as learning without a teacher. Self-Organizing Map (SOM)and outlier detection, Hierarchical clustering, and Cluster Analysis (K-Means clustering, Fuzzy clustering) are various unsupervised learning algorithms. [3][2]
3) *Reinforcement Learning:* In reinforcement learning, the model is trained to make a sequence of decisions. The goal is achieved in an uncertain and potentially complex manner. The model performs trial and error to bring up a solution to the problem. Deep Q Network, Q-Learning, State Action-Reward-State-Action, Deep Deterministic Policy Gradient (DDPG) are various reinforcement learning algorithms. [3]

VI. DATA REDUCTION METHOD OF IDS

Majority of Machine learning and data mining approaches couldn't work well with intrusion detection because of gigantic complexity and size of datasets. These techniques take huge computational time to classify attacks which makes implementation more difficult in real time environments. This is because of huge number of features are contained in network data which is to be processed by Intrusion Detection System. For better classification, quantity and quality of features matter and it helps us understand their importance and their correlation. If features selected are very less, then classification quality will reduce and if they are more than required, it will make loss of generalization. Experimental results show that accuracy and computational cost is improved when we use feature extraction techniques in Intrusion Detection Systems. Thus, dimensionality and feature reduction strategies are being used as a pre-processing step to improve accuracy and to reduce time for attack detection.



6.1 Block Diagram

A. *Feature Selection*

Feature selection techniques are used to discover subset of finest features which could improve overall outcome of the procedure and generate few errors. Another goal is to decrease computation time and storage utilization. In IDS, feature selection techniques are utilized to improve accuracy of attack detection. Some of the dominant feature selection methods are Principal component analysis (PCA), Information gain (IG), and genetic algorithm. Filter and Wrapper are two kind of features selection strategies in which we incorporate different Feature Selection methods.

1) In Wrapper technique, a classifier is utilized as a black box for evaluating optimal features. Such methods achieve great speculation, yet sometimes endure high dimensionality due to the computational expense of preparing the classifier.
2) Filter methods don't utilize any classifier for feature evaluation and are relatively powerful against overfitting, yet it utilizes autonomous estimation techniques, for example, distance measures, consistency measures, and correlation measure.

B. *Feature Extraction*

In dataset rows represent samples and columns represents features, where features are a result of quantitative and subjective findings. Feature extraction is used to reduce the dimensionality of data set by reducing set of features in a way that accuracy of attack detection is not altered and time used in discovery is reduced. Numerous feature extraction methods are available in the field.

C. *Clustering*

In clustering data samples are grouped into sets of data where data samples in each set is similar in one way or the other. [10]

VII. PERFORMANCE MATRIX

A. *Confusion Matrix*

Confusion matrices are used to represent the data associated to predicted and actual classification done by classifiers. Following terms are used while representing a confusion matrix.

1) *True-Positive (TP):* Correctly classify an anomalous sample as attack, here expected and actual values are identical, But the model anticipated a Positive value and the actual values is Positive.
2) *True-Negative (TN):* Correctly classify a non-attack sample as ordinary instance, here expected and actual values are identical but, model anticipated a Negative Values and the actual values is negative.
3) *False-Positive (FP):* Incorrectly classify an ordinary sample as anomalous instance, here predicted values turned out to be negative, in this model the actual values are negative and its expected as positive.
4) *False-Negative (FN):* Incorrectly classify an attack sample as ordinary instance.

Reduction of False negatives and False positives is a major research problem as these have very negative effects on overall security of networks. It is a matrix consisting of four possibilities, as shown in TABLE II, and through this matrix, it is possible to know the number of correctly classified records and the number of incorrectly classified records. [11]

Table 7.1. Performance Matrix

Confusion matrix		predict	
		Normal	Attack
Actual	Normal	True Negative (TN)	False Positive (FP)
	Attack	False Negative (FN)	True Positive (TP)

VIII. COMPARISON RELATED WORK

Table 8.1. Comparison related work

TITLE	ALGORITHM	DATASET	RESULT (ACCURACY)	FINDING	DRAWBACK
IDS using bagging with partial decision tree base classifier[12]	1)Genetic Algorithm (GA) based feature selection. 2)Bagged Classifier with partial decision tree	NLS-KDD99	Bagged Naïve Bayes=89.4882% Naïve Bays=89.6002% PART=99.6991% C4.5=99.6634% Bagged C4.5= 99.7158% Bagged PART=99.7166%	Reduced high false alarm	High time was required to build the model
IDS based on combining cluster centers and nearest neighbors[13]	1) k-Nearest Neighbor (k-NN) 2) Cluster Center and Nearest Neighbor 3) Support Vector Machine	KDD-Cup99	CANN=99.76% KNN=93.87% SVM=80.65%	Feature representation was applied for normal connections and attacks	U2L and R2L attacks were not effectively detected by CANN

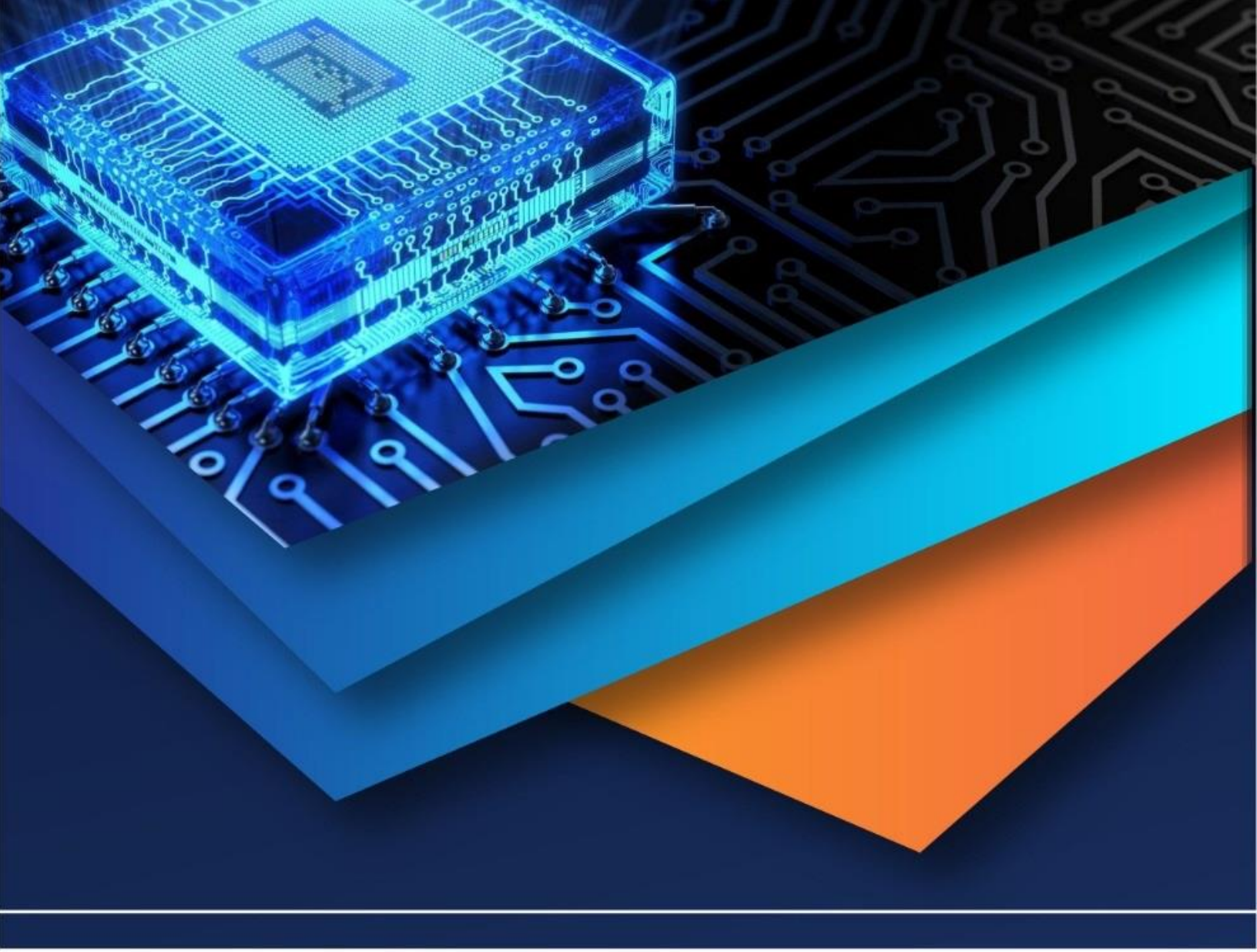
Comparison of classification techniques applied for network intrusion detection and classification[14]	1) Breadth-Forest Tree (BFTree) 2) Naïve Bayes Decision Tree (NBTree) 3) J48 4) Random Forest Tree (RFT) 5) Multi-Layer Perceptron (MLP) 6) Naïve Bayes	NSL-KDD	BFTree=98.24% NBTree=98.44% J48=97.68% RFT=98.34% MLP=98.53% NB=84.75%	Achieved reduction in false positive	There is need to evaluate the model on the most updated datasets.
Performance Comparison of IDS using Three Different ML Algorithm[15]	1) Support Vector Machine (SVM). 2)Random Forest(RF). 3)K-Nearest Neighbour(KNN).	NSL-KDD	SVM=99.87% RF=96.73% SVM=96.84%	The model is efficient as it returns a low false alarm and high detection rate in 13(FS)	A feature selection method like evolutionary computation needs to be applied to improve accuracy
Implementation of Machine Learning Algorithms for Detection of Network Intrusion[16]	1)Decision Tree (DT) 2) Logistic Regression (LR) 3) Random Forest (RF) 4) Support Vector Machine (SVM)	NSL-KDD	DT=72=303% LR=68.674% RF=73.784% SVM=71.779%	Showed that working with random forest in building IDS saves execution time	The model performs efficiently only with single classifier.
Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques[17]	1) Random Forest (RF) 2) Decision Tree (DT)	NSL-KDD	RF=95.323% DT=81.868%	Easily implemented	Slow detection rate and high false positive.
CA Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset[18]	1) Support Vector Machine(SVM) 2) Decision Tree (DT) 3) Random Forest Tree (RFT) 4) Logistic Regression(LR)	NSL-KDD	SVM=87.65% DT=99.21% RFT=99.48% LR=86.66%	Promising results for Dos Attack using different attributes.	Least accurate results for U2R.

IX. CONCLUSION

The introduction of ML invents new approaches for IDS whereby various researchers and academics have implemented diverse forms of classifications in the development of models of IDSs. The paper addressed numerous research papers written from 2015 to 2020 on the use of machine learning classifiers in intrusion detection systems. Ensemble and hybrid classifiers have been able to outperform their single classifier equivalent among the different models implemented in the various works being done, and thus have the highest predictive accuracy and detection rate.

REFERENCES

[1] M. Alkasassbeh and M. Almseidin, "Machine Learning Methods for Network Intrusion Detection." World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:12, No:8, 2018.
[2] Usman Shuiabu, Musu, Sudesha Chakraboty," A Review on Intrusion Detection System Using Machine Learning Techniques".2021 International conference on computing, Communication & Intelligent System vol:12, No:9, 2018
[3] Prachiti Parkar's "A Survey On Cyber Security IDS Using ML Methods". 2021 5th International Confereence on Intelligent company and control system. CFP21K74-ART; ISBN: 978-0-7381-1327-2
[4] Ripon Patigiri, Udit Varshey, Tanya Akutato & Rajesh Kunde "An Investiogation on Intrusion Detection Sstem Using Machine learning.
[5] Ali Foroughfar, Mohammad Abadeh, A momenzadeh, Maziyar Baran Pollyan "Misuse Detection Via a Novel Hybrid System".2009 Third UK Aim European Symposium.
[6] Dr.Rohini Nagapadma " Intrusion Detection System Using Naïve Byes Algoritm".2019 5th IEEE International WIE Conference on Electrical and Computer Enginerring
[7] Mr Subash Waskle, Mr Lokesh Parshar, Mr Upendra Singh "Intrusion Detection System Using PCA with Radom Forest."2020 International conference
[8] Rajesh Thomas, Deepa Pavithran" A Survey of Intrusion Detection Models based on NSL-KDD set".
[9] Gaurav Meena, Raj Choudhary "A Revew On IDS Classification Using KDD 99 and NSL KDD Dataset".
[10] Kunal and Mohit Dua, "Machine Learning Approach to IDS: A Comprehensive Review,". ISBN:978-1-7281-0167-5,DOI: 10.1109/ICECA.2019.882212010
[11] Hamza Nachan , Pratik Kumhar , Simran Birla , Dristi Poddar, Sambhaji Sarode, " Intrusion Detection System: A Survey,". Volume 10, Issue05(May2021),DOI:10.17577/IJERTV10IS050479
[12] D. P. Gaikwad and R. C. Thool, "Intrusion detection system using Bagging with Partial Decision Tree base classifier," Procedia Comput. Sci., vol. 49, no. 1, pp. 92–98, 2015, doi: 10.1016/j.procs.2015.04.231.
[13] W. C. Lin, S. W. Ke, and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowledge-Based Syst., vol. 78, no.1,pp.13–21,2015,doi: 10.1016/j.knosys.2015.01.009.
[14] A. S. Amira, S. E. O. Hanafi, and A. E. Hassanien,"Comparison of classification techniques applied for network intrusion detection and classification," J. Appl. Log., vol. 24,pp.109–118,2017,doi: 10.1016/j.jal.2016.11.018.
[15] Zena Khalid Ibrahim, Mohammed Younis Thanon, "Performance Comparison of Intrusion Detection System Using Three Different Machine Learning Algorithms,". ISBN: 978-1-7281-8501-9, Doj: 10.1109\ICIT50816.2021
[16] M. Sazzadul Hoque, "An Implementation of Intrusion Detection System Using Genetic Algorithm,". Netw. Secur. Its Appl., vol. 4, no. 2,pp.109120,2012,doi:10.5121/ijnsa2012.4208.
[17] Bhavani T. T, Kameswara M. R and Manohar A. R. (2020). Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques. International Conference on Sustainable Technologies for Computational Intelligence (ICSTCI). (pp. 637-643). Springer



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY